UET@eRisk2025: Severity Estimation for Depression Symptoms Searching and Early Risk Detection

Notebook for the eRisk Lab at CLEF 2025

Tu-Phuong Mai[†], Minh-Ha H. Le[†], Duc-Luong Tran[†], Duy-Cat Can and Hoang-Quynh Le[*]

VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

Abstract

In this working note, we describe our participation in Task 1 and Task 2 of the CLEF eRisk 2025 Lab, which focuses on the early detection of depression based on Reddit user-generated content. For Task 1, which involves ranking up to 1,000 sentences according to their relevance to each of the 21 BDI-II depressive symptoms, we combined symptom classification with two approaches: (i) semantic similarity-based, where clustering techniques to group and rank sentences based on their relevance to specific depressive symptoms; and (ii) machine learning-based, where we use the output scores from a model fine-tuned for symptom detection and directly rank sentences based on predicted relevance scores. For Task 2, which targets early detection of depression within multi-user conversations, we design a multi-stage architecture that performs sentence-level symptom and severity detection, aggregates these signals at the post level, and finally estimates depression risk at the conversation level. This layered structure allows the model to capture both localized symptom cues and broader conversational patterns.

Keywords

Depression, Symptoms Searching, Early Risk Detection, Natural Language Processing, Social Media

1. Introduction

The CLEF eRisk 2025 Lab [1, 2] focuses on the early detection of mental health risks through the analysis of online user-generated content. Competition promotes the development of natural language processing (NLP) systems that are capable of identifying early signs of depression based on social media text. The data used in eRisk tasks is collected from the Reddit platform, where users share personal experiences through posts or discussions. These environments often encourage openness and anonymity, resulting in large volumes of natural language data that reflect individuals' thoughts, emotions, and behaviors. This year, eRisk 2025 features three tasks: (1) sentence ranking for depression symptoms, based on the 21 symptoms from the Beck Depression Inventory-II (BDI-II) questionnaire [3]; (2) contextualized early detection of depression, using full, multi-user conversational threads presented in chronological order; and (3) a pilot task involving the detection of depression in LLM-powered conversational agents, where systems must infer the mental state of a simulated user. Together, these tasks aim to support the development of practical and scalable methods for mental health monitoring and early intervention.

Task 1 continues the setup from the eRisk 2024 challenge [1], several teams employed retrieval-based approaches by ranking user-generated content based on its cosine similarity to the Beck Depression Inventory-II (BDI-II) questionnaire [3]. Among them, the NUS-IDS team [4] achieved top performance by leveraging ensemble learning and contrastive fine-tuning. Their system combined sentence-transformer models fine-tuned on task-specific data with expressive exemplars generated via prompting GPT-4 [5], incorporating both BDI symptoms and features from the Early Maladaptive Schemas (EMS) taxonomy [6]. Task 2 continues the setup from eRisk 2022 [7], where the NLPGroup-IISERB team [8] attained top performance using entropy-based bag-of-words features combined with an SVM classifier.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*]Corresponding author.

[†]These authors contributed equally.

21020552@vnu.edu.vn (T. Mai); 21020621@vnu.edu.vn (M. H. Le); 22021148@vnu.edu.vn (D. Tran); catcd@vnu.edu.vn (D. Can); lhquynh@vnu.edu.vn (H. Le)

Their approach demonstrated that traditional feature engineering, when carefully designed, can remain competitive for early risk detection.

Our team participated in Task 1 and Task 2 of the CLEF eRisk 2025 Lab. We leverage DepRoBERTa [9], a RoBERTa-based [10] model pre-trained for depression detection, in both tasks to filter out irrelevant sentences that do not reflect depressive content. In Task 1, after identifying relevant sentences using the filtering model, we adopt two approaches: (i) a semantic similarity-based method, where we cluster sentence embeddings to group semantically similar expressions for each symptom and rank sentences based on their distance to cluster centroids; and (ii) a machine learning-based method, where we use the output scores from a multi-task DepRoBERTa model fine-tuned for symptom detection and directly rank sentences based on predicted relevance scores. For Task 2, we proposed a multi-stage framework that first used the same filtering model, which was further utilized to produce sentence-level embeddings, which were used to detect symptom presence and estimate severity. Then aggregates this information at the post and conversation levels to estimate depression risk, integrating both local and contextual cues for early mental health detection.

2. Dataset

2.1. Task 1: Search for Symptoms of Depression

This task focuses on ranking documents relevant to symptoms of depression as outlined in the BDI-II questionnaire. The goal is to produce ranked lists containing up to 1,000 of the most relevant sentences for each specific symptom. Evaluation involved expert annotators labeling pooled candidate sentences as relevant if they addressed the symptom and reflected the individual's state, with context provided for accuracy. The final relevance scores were determined using two approaches: majority voting, where a sentence is marked relevant if most assessors agree, and unanimity, where all assessors must agree on relevance. These methods ensure reliable and consistent evaluation for training and testing.

The training data for Task 1 was provided from previous editions of the same task, specifically from eRisk 2023 and eRisk 2024. The test set for this year includes data collected from 9,000 Reddit users, comprising over 17 million sentences. The data is formatted according to the TREC format. The main statistics[1] of the corpus are presented in Table 1.

Table 1
Corpus statistics for Task 1: Search for Symptoms of Depression.

	Training set	Testing set
Number of users	554,418	9,000
Number of sentences	19,349,315	17,553,441
Average number of words per sentence	17.12	12.39

2.2. Task 2: Contextualized Early Detection of Depression

This task focuses on the early detection of depression by analyzing full conversational contexts. Unlike previous tasks that consider isolated user posts, this task processes interactions among all participants in a conversation sequentially, reflecting real-world social media dynamics. The dataset includes the target user's writing history and all comments from conversation members, enabling timely depression detection based on evolving dialogue.

The dataset follows the format described in Losada & Crestani (2016)[11] and consists of Reddit conversations where each conversation forms a tree-structured thread centered around a target user. The objective is to predict a depression score $y \in [0, 1]$ for the target user based on contextual signals from the conversation.

[1]Statistics of the training set are based on reports from the eRisk 2023 and eRisk 2024.

3. Proposed Method

3.1. Task 1: Search for Symptoms of Depression

Our approach to Task 1 is based on two directions: (i) a semantic similarity-based ranking pipeline and (ii) a machine learning-based ranking model.

Semantic similarity-based approach. This direction first uses a multi-label classification model to filter out irrelevant sentences. The remaining relevant sentences are embedded using sentence transformers and grouped into symptom-specific clusters. At inference, test sentences are ranked based on their similarity to these clusters. We explore three configurations: (a) direct semantic similarity, (b) embeddings fine-tuned via contrastive learning, and (c) an ensemble of multiple embedding models.

Machine learning-based approach. In this direction, we directly use the output scores from a fine-tuned multi-task model (described in Section 3.1.3) to rank sentences by relevance.

3.1.1. Pre-processing

We began with the official sentence-level annotations provided in TREC format, where each sentence is associated with a user ID and timestamp. To ensure high-quality input and reduce noise, we applied several filtering steps. Texts were lowercased, and non-linguistic tokens such as URLs, emojis, and special characters were removed. Crucially, we filtered for first-person expressions by detecting first-person pronouns (e.g., "I", "me", "my", ...), under the hypothesis that self-reported experiences better reflect the user's mental state than statements about others or general opinions. The resulting dataset included relevant sentences from the 2023 and 2024 editions of eRisk, which were used for model training and clustering.

3.1.2. Semantic similarity-based approach

We combine filtering with clustering to identify semantically representative symptom expressions. First, sentences are filtered using a DepRoBERTa-based multi-label classifier to retain only those relevant to any of the 21 BDI-II symptoms. After filtering for relevant sentences, we group them into semantic clusters and rank new sentences based on their distance to these the nearest centroid. This enables the system to identify symptom-relevant sentences that may express depressive cues in more varied ways.

Clustering and Semantic Representation. To capture variations in how each symptom is linguistically expressed, we performed clustering over the relevant training sentences. Each sentence s_i was embedded using a Sentence Transformer [12] model - specifically the `nomic-embed-text-v1.5` [13] - to obtain a $d - dimensional$ vector representation:

$$\mathbf{v}_i = \text{Embed}(s_i) \in \mathbb{R}^d \tag{1}$$

For each symptom $c \in \{1, \ldots, C\}$ (with $C = 21$ for BDI-II symptoms), we collected the subset of training embeddings $\{\mathbf{v}_i^{(c)}\}$ relevant to that symptom. Then, we applied K-means clustering to this set to form k clusters:

$$\{\boldsymbol{\mu}_1^{(c)}, \ldots, \boldsymbol{\mu}_k^{(c)}\} = \text{KMeans}\left(\{\mathbf{v}_i^{(c)}\}\right) \tag{2}$$

where $\boldsymbol{\mu}_j^{(c)}$ denotes the centroid of the j-th cluster for symptom c.

This clustering strategy groups semantically similar sentences into coherent sub-themes within each symptom category. The choice of k could be made to balance between intra-cluster similarity and inter-cluster diversity.

Contrastive Learning. To improve the discriminative quality of sentence embeddings, we applied contrastive learning using the InfoNCE loss. Each sentence embedding $\mathbf{v}_i$ obtained from the `nomic-embed-text-v1.5` model was first projected into a lower-dimensional space via a linear mapping layer:

$$\mathbf{h}_i = W \cdot \mathbf{v}_i + \mathbf{b}, \quad \mathbf{h}_i \in \mathbb{R}^{128} \tag{3}$$

where $W \in \mathbb{R}^{128 \times d}$ is a trainable weight matrix and d is the original embedding dimension.

Given a batch of training samples with known symptom labels, positive pairs were constructed from sentences annotated with the same symptom, and negatives from sentences belonging to different symptoms. The InfoNCE loss was then applied to pull embeddings of similar sentences closer and push dissimilar ones apart:

$$\mathcal{L}_{h_i} = -\log \frac{\sum_{p \in P(i)} \exp(\text{sim}(\mathbf{h_i}, \mathbf{h_p})/\tau)}{\sum_{a \in A(i)} \exp(\text{sim}(\mathbf{h_i}, \mathbf{h_a})/\tau)} \tag{4}$$

Where:

- $P(i) = \{j \neq i | y_j = y_i\}$: Set of positive indices with the same label as anchor.
- $A(i) = \{j \neq i\}$: Set or all samples.
- $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature hyperparameter.

This training objective encourages the embedding space to reflect symptom-level semantic distinctions more clearly, enhancing the quality of downstream clustering and similarity-based ranking.

Sentence Assignment and Ranking. Each test sentence predicted as relevant was also embedded using the same nomic model. Then, for each symptom, we applied k-nearest neighbor search ($k = 11$) to identify the closest cluster centroid (among training clusters of that symptom). We assigned each test sentence to the nearest cluster and computed its distance to the centroid. This distance was converted to a normalized similarity score via:

$$\text{Similarity}(s_i) = 1 - \frac{\|\mathbf{v}_i - \boldsymbol{\mu}_i^c\|_2}{\max_{j \in \mathcal{S}_c}(\|\mathbf{v}_j - \boldsymbol{\mu}_j^c\|_2)} \tag{5}$$

Where:

- $\mathbf{v}_i$ is the embedding vector of test sentence s_i.
- $\boldsymbol{\mu}_i^c$ is the nearest cluster centroid for symptom c.
- $\mathcal{S}_c$ is the set of all test sentences predicted as relevant to symptom c.

The final ranking was derived by sorting all test sentences for each symptom in descending order of similarity, selecting the top 1,000 as the system output.

This approach combines high-precision filtering from the symptom classifier with semantic granularity from clustering, enabling the system to surface sentences that are not only relevant to a symptom but also representative of its most prototypical or central expressions.

3.1.3. Machine Learning-based approach

Given that the severity of a sentence often correlates with the presence and intensity of specific depressive symptoms, we adopt a multi-task learning approach to jointly model both aspects. Specifically, we fine-tune a DepRoBERTa [9] model to simultaneously predict symptom presence (as a 21-dimensional multi-label output) and estimate severity (as a continuous score in $[0, 1]$). This joint training not only allows the two tasks to benefit from shared representations but also encourages the model to capture subtle linguistic cues that reflect both the type and intensity of depressive expressions. This model takes an individual sentence as input and produces two outputs: a binary vector indicating the presence of relevant symptoms, and a scalar severity score.

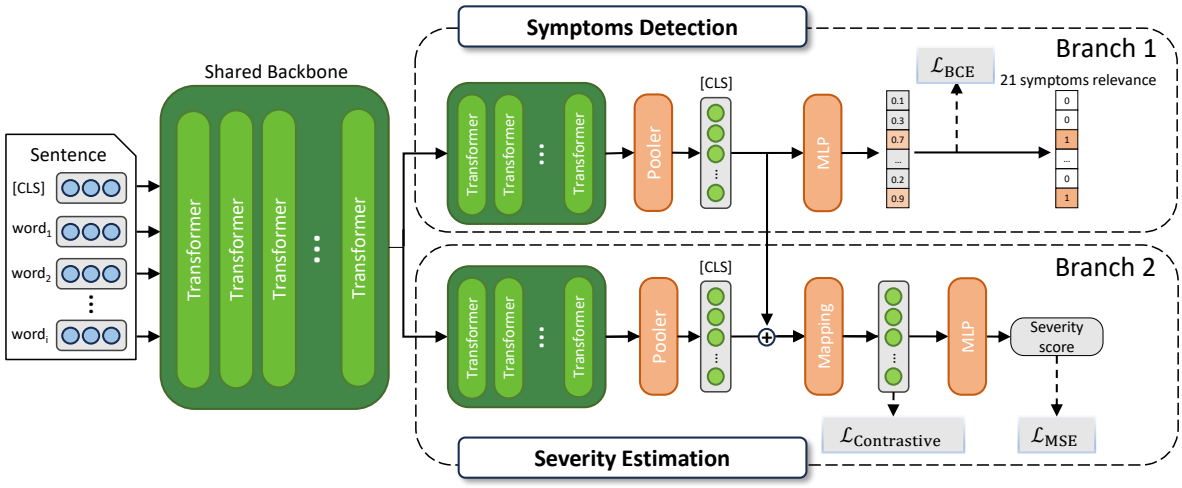

Figure 1: Sentence-level Symptom Detection and Severity Estimation model

Severity Label Generation Using Large Language Models. To create a reliable dataset for sentence-level symptom detection and severity estimation, we extended the Task 1 training data, which includes annotations for symptom relevance but lacks severity labels, by generating severity scores using a large language model (LLM). For each relevant sentence, we prompted the LLM with the sentence text and corresponding BDI-II symptom descriptions to assign a severity score in $\{0, 1, 2, 3\}$ based on the BDI-II criteria. These scores were then normalized to a continuous scale in $[0, 1]$. This process leverages both the clinical structure of BDI-II and the contextual reasoning capabilities of the LLM to provide consistent and meaningful severity annotations. The resulting dataset, containing both relevance and severity scores, enables supervised training of a multi-task model while avoiding the need for costly manual labeling.

Architecture. Figure 1 illustrates the architecture of our multi-task fine-tuned DepRoBERTa model for sentence-level symptoms detection and severity estimation. The architecture consists of:

- **Shared Backbone:** The first 18 layers of the pre-trained DepRoBERTa model are frozen during training.
- **Branch 1 — Symptom Detection:** A task-specific branch with 6 transformer layers and a pooler, followed by a multi-label classification head to predict the relevance of a sentence to 21 depression-related symptoms.
- **Branch 2 — Severity Estimation:** Another 6-layer branch with a pooler. The pooled vector from this branch is concatenated with the pooled output from Branch 1, then passed through a linear mapping layer to combine features. The resulting vector is used both for computing contrastive loss and as input to a regression MLP head that outputs the severity score $s \in [0, 1]$.

Training Strategy. We employ a two-phase training procedure:

1. **Phase 1:** Train the symptom detection branch while freezing the severity estimation branch.
2. **Phase 2:** Once Branch 1 stabilizes, we freeze it and start training Branch 2.

Loss Functions. The model is optimized using a combination of three loss components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{MSE}} + \lambda \cdot \mathcal{L}_{\text{InfoCL}} \tag{6}$$

where:

- $\mathcal{L}_{\text{BCE}}$: Binary cross-entropy loss for multi-label classification.
- $\mathcal{L}_{\text{MSE}}$: Mean squared error for severity score regression.
- $\mathcal{L}_{\text{InfoCL}}$: Contrastive loss (InfoNCE [14]) applied on the pooled sentence embeddings from Branch 2 to improve representation quality.
- λ: A weighting factor to balance the contrastive loss.

Sentence Ranking Given a post or comment p consisting of n sentences: $p = \{s_1, s_2, \ldots, s_n\}$, each sentence s_i is passed through a multi-task model:

$$f_{\text{sym}}(s_i) = \mathbf{z}_i \in \{0,1\}^{21}, \quad f_{\text{sev}}(s_i) = r_i \in [0,1] \tag{7}$$

where $\mathbf{z}_i$ is a binary symptom vector for the 21 depressive symptoms, and r_i is the predicted severity score if s_i is relevant.

For each symptom $c \in \{1, \ldots, 21\}$, we rank all sentences by their predicted probability $z_i[c]$ in descending order and select the top 1000 sentences. This method directly uses the model's outputs to perform sentence ranking and was used in our best-performing configuration (Run 4).

3.1.4. Submitted Configurations

We submitted five configurations for Task 1, described as follows:

Run 0: Similarity. Semantic similarity-based approach using the original `nomic-embed-text-v1.5` model without contrastive learning. K-means clustering was applied with $k = 11$ to form symptom-specific clusters.

Run 1: Ensemble Similarity. An ensemble of three cluster-based similarity runs: (i) `nomic-embed-text-v1.5` with $k = 5$, (ii) `nomic-embed-text-v1.5` with $k = 11$, and (iii) `modernbert-embed-base` with $k = 11$.

Run 2: Contrastive Learning Similar to Run 0 but using contrastive learning, fine-tuned `nomic-embed-text-v1.5` embeddings. Embeddings were projected to a 128-dimensional space and trained using InfoNCE loss to improve symptom-level semantic separation.

Run 3: Ensemble Contrastive Learning. An ensemble combining Run 1 and Run 2, leveraging both diverse embedding sources and contrastive learning enhanced representations for more robust similarity ranking.

Run 4: Machine Learning. A machine learning-based approach using output scores from the fine-tuned multi-task model described in Section 3.1.3. This model directly predicts symptom relevance and severity, and its scores are used to rank the sentences.

3.2. Task 2: Contextualized Early Detection of Depression

Our pipeline consists of three stages:

1. **Sentence-level symptom detection and severity estimation**: Each sentence is analyzed to identify the presence of depressive symptoms and to assign a fine-grained severity score. This stage uses a multi-task model, which is described in Section 3.1.3.
2. **Post-level depression scoring**: Relevant sentence representations and their associated severity scores are aggregated to compute a depression score for each post or comment.
3. **User-level depression estimation**: Finally, a set of rule-based heuristics is applied to combine post-level scores across the conversation tree, yielding a final depression score for the target user.

3.2.1. Pre-processing

We adopt the dataset released by the organizers, which filters out noisy or incomplete threads. To preserve relevant contextual information, each conversation tree is pruned to keep only the branches:

- leads to the target user (i.e., ancestor nodes),
- or are direct responses from the target user (i.e., children nodes).

In cases where parent nodes are missing, dummy nodes are inserted to maintain the tree structure and avoid losing conversation branches.

3.2.2. Post-level Depression Scoring

In this stage, we aggregate sentence-level information to estimate a depression score for each post or comment. Let $\{s_j\}_{j=1}^{m}$ denote the set of relevant sentences extracted from a given post, each associated with a severity score $r_j \in [0, 1]$.

Each sentence s_j is encoded using the fine-tuned DepRoBERTa model and extracts the sentence representation from the Pooler layer of Branch 2.

$$\mathbf{h}_j = \text{Pooler}(\text{DepRoBERTa}(s_j)) \tag{8}$$

The sequence of sentence embeddings $\{\mathbf{h}_j\}_{j=1}^{m}$ is passed through a bidirectional LSTM to capture contextual dependencies among sentences:

$$\mathbf{h}^{\text{text}} = \text{BiLSTM}_{\text{text}}(\{\mathbf{h}_j\}_{j=1}^{m}) \in \mathbb{R}^d \tag{9}$$

Similarly, the sequence of scalar severity scores $\{r_j\}_{j=1}^{m}$ is fed into a separate BiLSTM to capture the temporal structure and progression of severity:

$$\mathbf{h}^{\text{sev}} = \text{BiLSTM}_{\text{sev}}(\{r_j\}_{j=1}^{m}) \in \mathbb{R}^d \tag{10}$$

The final representation is obtained by concatenating the textual and severity embeddings, followed by a multi-layer perceptron (MLP) with a sigmoid activation to produce the depression score $\hat{y} \in [0, 1]$:

$$\hat{y} = \sigma(\text{MLP}([\mathbf{h}^{\text{text}}; \mathbf{h}^{\text{sev}}])) \tag{11}$$

3.2.3. User-level Depression Prediction

In the final stage, we aggregate sentence-level severity scores to produce a depression prediction for the target user. Given the severity scores of relevant posts across a conversation tree, we implement multiple rule-based configurations to explore different aggregation strategies. Each configuration defines specific rules for decision making:

Run 0: Target Node Only, Max Score We consider only the posts authored by the target user (target nodes) and take the maximum severity score as the final prediction:

$$\hat{y} = \max_{p \in P_{\text{target}}} (\text{score}(p)) \tag{12}$$

Where:

- P_{target} is the set of posts authored by the target user in the current conversation.
- $\text{score}(p)$ denotes the predicted severity score (in $[0, 1]$) for post p.

Run 1: Temporal Accumulation, Target Nodes Only We include historical posts of the target user and again use the maximum score across all such posts:

$$\hat{y} = \max_{p \in \{P_{\text{ht}} \cup P_{\text{ct}}\}} (\text{score}(p)) \tag{13}$$

where:

- P_{ct} is the set of posts authored by the target user in the current conversation.
- P_{ht} is the set of posts authored by the same user in previous conversations.
- $\text{score}(p)$ denotes the predicted severity score (in $[0, 1]$) for post p.

Run 2: Temporal Accumulation with Bonus Similar to run 1, but we add a bonus score if the maximum severity exceeds a high threshold τ_{high}. The final score is computed as:

$$\hat{y} = \max_{p \in \{P_{\text{ht}} \cup P_{\text{ct}}\}} (\text{score}(p)) + k_{\text{bonus}} \cdot \mathbb{L}\left[\max_{p \in \{P_{\text{ht}} \cup P_{\text{ct}}\}} (\text{score}(p)) > \tau_{\text{high}}\right] \tag{14}$$

Where:

- P_{ct} is the set of posts authored by the target user in the current conversation.
- P_{ht} is the set of posts authored by the same user in previous conversations.
- $\text{score}(p)$ is the predicted severity score (in $[0, 1]$) for post p.
- τ_{high} is the high depression score threshold.
- k_{bonus} is the bonus term added to the final score when the threshold condition is met.
- $\mathbb{L}[\cdot]$ is the indicator function that returns 1 if the condition inside is true, otherwise 0.

Run 3: Temporal Accumulation with Neighbor-based Uncertainty Handling We consider both current and historical posts from the target user. For each post $p \in \mathcal{P}_{\text{target}}$, if its severity score falls within an uncertainty range $[\tau_{\text{low}}, \tau_{\text{high}}]$, we apply a neighbor-based adjustment using its parent and root scores. The adjusted score s'_p is defined as:

$$s'_p = \begin{cases} (1-\alpha) \cdot s_p + \alpha \cdot s_{\text{parent}(p)} + \beta \cdot s_{\text{root}} & \text{if } \tau_{\text{low}} < s_p \leq \tau_{\text{high}} \\ s_p & \text{otherwise} \end{cases} \tag{15}$$

where:

- s_p is the original severity score of post p.
- $s_{\text{parent}(p)}$ is the score of the parent node of p in the conversation tree.
- s_{root} is the score of the root node of the conversation.
- τ_{low} is the low depression score threshold.
- τ_{high} is the high depression score threshold.
- α is the weight for the parent node influence.
- β is the weight for the root node influence.

The final decision score is the maximum of all adjusted scores:

$$\hat{y} = \max_{p \in \mathcal{P}_{\text{target}}} (s'_p) \tag{16}$$

Run 4: Temporal Accumulation with Community-based Adjustment We accumulate both current and historical posts from the target user. For each target post $p \in \mathcal{P}_{\text{all target}}$, we consider all posts in the conversation branch from the root node to p, excluding p itself, as:

$$\mathcal{C}_p = \{q \in \text{Branch}(r, p) \mid q \neq p\} \tag{17}$$

Let s_p be the original severity score of p, and $s_{\mathcal{C}_p}$ be the average severity score of the community:

$$s_{\mathcal{C}_p} = \frac{1}{|\mathcal{C}_p|} \sum_{q \in \mathcal{C}_p} s_q \tag{18}$$

$$s'_p = \begin{cases} s_p + k_{\text{bonus}} \cdot (s_{\mathcal{C}_p} - s_p) & \text{if } s_{\mathcal{C}_p} > \max(\tau_{\text{high}}, s_p) \\ s_p + k_{\text{penalty}} \cdot (s_{\mathcal{C}_p} - s_p) & \text{if } s_{\mathcal{C}_p} < \min(\tau_{\text{low}}, s_p) \\ s_p & \text{otherwise} \end{cases} \tag{19}$$

The final prediction score is the maximum over all adjusted scores:

$$\hat{y} = \max_{p \in \mathcal{P}_{\text{target}}} (s'_p) \tag{20}$$

Where:

- s_p is the original severity score of post p.
- τ_{high}, τ_{low} are the high depression score and low depression score threshold.
- k_{bonus}, k_{penalty} are the bonus term added to the final score and the penalty term subtracted from the final score when the threshold condition is met.

Each run serves as a configuration of the decision logic and can be evaluated independently to assess the robustness of rule-based aggregation methods over tree-structured social media conversations.

4. Evaluation Results & Discussion

4.1. Task 1: Search for Symptoms of Depression

A total of 67 runs from participants were submitted for this task. In Table 2, we present the ranking-based evaluation results for Task 1 (majority setting), comparing the best configuration from each participating team. Our submission achieved the second-best performance in both NDCG and AP, while also maintaining strong results across R-PREC and P@10. This demonstrates that our approach offers a well-balanced trade-off between ranking quality and precision.

Tables 3 show the ranking-based performance of our system under majority voting schemes. Among our runs, the *machine learning* configuration consistently achieves the best results, notably with an NDCG of 0.623 (majority) and 0.577 (unanimity), highlighting its effectiveness. Similarity-based approaches also perform reasonably well, with slight improvements when ensembling is applied. In contrast, contrastive learning methods underperform across all metrics, suggesting they may not be well-suited for this task without further tuning.

4.2. Task 2: Contextualized Early Detection of Depression

We report the results of our models on the public leaderboard in Table 4. Among our runs, *Run 2: Temporal Accumulation with Bonus* consistently yields the best performance, with an F_{latency} of 0.68 and F_1 of 0.73, demonstrating the benefit of incorporating historical context and severity-based reward.

A total of 50 runs from 12 participants were submitted for this task. In Table 5, we present the decision-based evaluation results for Task 2, comparing the best configuration from each participating team. Our submission achieved the fourth-best performance in both $F1$ and F_{latency}.

Table 2
Ranking-based evaluation for Task 1 (majority) of the best configuration of all teams.

Team	Run	AP	R-PREC	P@10	NDCG
INESC-ID	unanimity	**0.354**	**0.433**	**0.876**	0.575
SonUIT	config4	0.328	<u>0.426</u>	0.767	0.578
PJs-team	teamRRens-v2	0.279	0.360	<u>0.800</u>	0.503
BGU-Data-Science	sbert	0.240	0.324	0.743	0.516
NYCUNLP	01	0.237	0.322	0.662	0.501
ixa_ave	base_filter30	0.102	0.203	0.338	0.342
LHS712-Team-1	BERT CONSENSUS	0.102	0.199	0.529	0.321
COMFOR	bert ranked	0.013	0.041	0.243	0.082
COTECMAR-UTB	ranked updated	0.077	0.165	0.414	0.290
ThinkIR	2025	0.068	0.157	0.409	0.228
ELiRF-UPV	model1	0.035	0.101	0.100	0.216
Team-Gryffindor	task1	0.017	0.042	0.019	0.183
RELAI	2	0.008	0.036	0.038	0.078
UniORNLP-dahlia	simple hyde	0.014	0.040	0.205	0.072
HULAT_UC3M	roberta	0.018	0.034	0.363	0.065
Synapse	HighestSimilarityFirst	0.001	0.002	0.038	0.009
UET-Psyche-Warriors	machine learning	<u>0.339</u>	0.394	0.776	<u>0.623</u>

*Best result of each metric is highlighted in **bold**.*
The second best result of each metric is highlighted in <u>underline</u>.

Table 3
Ranking-based evaluation for Task 1 (majority) of our configurations.

Team	Run	AP	R-PREC	P@10	NDCG
UET-Psyche-Warriors	similarity (0)	0.311	0.378	0.657	0.588
UET-Psyche-Warriors	ensemble similarity (1)	0.315	0.390	0.657	0.612
UET-Psyche-Warriors	contrastive learning (2)	0.165	0.258	0.457	0.450
UET-Psyche-Warriors	ensemble contrastive learning (3)	0.147	0.228	0.462	0.419
UET-Psyche-Warriors	machine learning* (4)	**0.339**	**0.394**	**0.776**	**0.623**

*Best configuration is highlighted by *.*
*Best result of each metric is highlighted in **bold**.*

Table 4
Decision-based evaluation for Task 2 of our configurations.

Team	Run	P	R	F1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	speed	$F_{latency}$
UET-Psyche-Warriors	0	**0.67**	0.78	0.72	0.10	0.06	24.50	0.91	0.66
UET-Psyche-Warriors	1	0.63	0.85	0.72	0.09	0.05	16.00	0.94	0.68
UET-Psyche-Warriors	2*	0.63	**0.86**	**0.73**	0.09	**0.04**	16.00	0.94	**0.68**
UET-Psyche-Warriors	3	0.63	0.85	0.72	0.09	0.05	16.00	0.94	0.68
UET-Psyche-Warriors	4	0.63	0.84	0.72	0.09	0.05	15.50	0.94	0.68

*Best configuration is highlighted by *.*
*Best result of each metric is highlighted in **bold**.*

5. Conclusion

In this report, we present our approaches for both Task 1 and Task 2 of the eRisk 2025 challenge, focusing on early detection and symptom identification of depression from social media posts. For Task 1, we explored various ranking-based methods based on two approaches: (i) Semantic similarity-based methods that cluster sentence embeddings and rank by proximity to symptom centroids, and (ii)

Table 5
Decision-based evaluation for Task 2 of top teams.

Team	Run	P	R	F1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	speed	$F_{latency}$
HIT-SCIR	4	0.77	0.94	**0.85**	0.09	**0.03**	8.00	0.97	**0.82**
ELiRF-UPV	0	**0.78**	0.81	0.79	0.08	0.04	7.00	0.98	0.78
HU	1	0.72	0.77	0.75	0.10	0.05	11.00	0.96	0.72
PJs-team	0	0.66	0.75	0.71	0.09	0.06	17.00	0.94	0.66
Lotu-Ixa	3	0.53	0.78	0.63	**0.05**	**0.03**	**1.00**	**1.00**	0.63
SINAI-UJA	0	0.24	**1.00**	0.39	0.08	0.05	3.00	0.99	0.38
NYCUNLP	4	0.20	0.93	0.33	0.16	0.07	18.00	0.93	0.31
COTECMAR-UTB	0	0.29	0.65	0.40	0.12	0.10	69.00	0.74	0.29
FU-TU-DFKI	0	0.17	0.97	0.29	0.16	0.07	11.00	0.96	0.28
Capy-team	3	0.11	**1.00**	0.20	0.11	0.10	**1.00**	**1.00**	0.20
DS-GT	0	0.11	**1.00**	0.20	0.12	0.10	2.00	**1.00**	0.20
UET-Psyche-Warriors	2	0.63	0.86	0.73	0.09	0.04	16.00	0.94	0.68

*Best result of each metric is highlighted in **bold**.*
The second best result of each metric is highlighted in underline.

Machine learning-based methods that directly use the output scores from the multi-task model. Among these, the second approach achieved the best performance across evaluation metrics, demonstrating the effectiveness of our fine-tuned multi-task model for sentence-level symptom detection.

For Task 2, we designed several temporal aggregation strategies to detect early warning signs of depression. These configurations leverage both current and historical user data, with enhancements such as uncertainty handling and community-based score adjustment. The most effective setup integrated severity scoring with threshold-based boosting, resulting in competitive latency-aware performance.

Across both tasks, we incorporated the multi-task model to sentence ranking in Task 1 and provided representations while filtering out irrelevant content in Task 2, contributing to improved robustness and precision. Overall, our approaches highlight the effectiveness of combining fine-tuned language models with task-specific heuristics and temporal context for early detection of mental health risks.

Declaration on Generative AI

In the preparation of this report, we only used Grammarly and ChatGPT for spell/grammar checking and improving the readability of the manuscript. No part of the content, analyses, or results was generated by AI tools. All methodological design, implementation, experiments, and interpretations were conducted solely by the authors.

References

[1] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, Part II, volume To be published of *Lecture Notes in Computer Science*, Springer, 2025.

[2] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9-12 September, 2025, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[3] A. T. Beck, Beck depression inventory–ii, Psychological assessment (1996).

[4] B. H. Ang, S. D. Gollapalli, S.-K. Ng, Nus-ids@ erisk2024: ranking sentences for depression symptoms using early maladaptive schemas and ensembles, Working Notes of CLEF (2024) 9–12.

[5] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[6] J. E. Young, Cognitive therapy for personality disorders: A schema-focused approach, Professional Resource Press/Professional Resource Exchange, 1999.

[7] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2022: pathological gambling, depression, and eating disorder challenges, in: European Conference on Information Retrieval, Springer, 2022, pp. 436–442.

[8] H. Srivastava, N. Lijin, S. Sruthi, T. Basu, Nlp-iiserb@ erisk2022: Exploring the potential of bag of words, document embeddings and transformer based framework for early prediction of eating disorder, depression and pathological gambling over social media., in: CLEF (Working Notes), 2022, pp. 972–986.

[9] R. Poświata, M. Perełkiewicz, OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 276–282. URL: https://aclanthology.org/2022.ltedi-1.40. doi:10.18653/v1/2022.ltedi-1.40.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[11] D. Losada, F. Crestani, A test collection for research on depression and language use, volume 9822, 2016, pp. 28–39. doi:10.1007/978-3-319-44564-9_3.

[12] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[13] Z. Nussbaum, J. X. Morris, B. Duderstadt, A. Mulyar, Nomic embed: Training a reproducible long context text embedder, 2025. URL: https://arxiv.org/abs/2402.01613. arXiv:2402.01613.

[14] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).